



デジタル通信とネットワーク

journal homepage: www.keaipublishing.com/dcan

6Gネットワークにおける擬似LiDAR信頼度サンプリングと階層的幾何学的特徴抽出による単眼3D物体検出

Jianlong Zhang^a, Guangzu Fang^a, Bin Wang^{a, **}, Xiaobo Zhou^b, Qingqi Pei^c, Chen Chen^{c,*}^a School of Electronic Engineering, Xidian University, Xi'an, 710071, China^b School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, 300000, China^c School of Telecommunications Engineering, Xidian University, Xi'an, 710071, China

ARTICLE INFO

Keywords:

Monocular 3D object detection

Pseudo-LiDAR

Confidence sampling

Hierarchical geometric feature extraction

ABSTRACT

6Gネットワーク技術の高帯域幅と低遅延により、車両プラットフォームへの単眼3D物体検出の適用を成功させることができる。単眼3D物体検出ベースの擬似LiDARは、自律走行分野におけるLiDARソリューションと比較して、低コストで低消費電力のソリューションである。しかし、この手法には、(1)単眼奥行き推定の非線形誤差分布に起因する生成擬似LiDAR点群の品質の低さ、(2)LiDARベースの3D検出ネットワークに存在する点群の無視されたグローバルな幾何学的構造特徴に起因する点群特徴の表現能力の低さ、などの問題がある。そこで、単眼3次元物体検出のための擬似LiDAR信頼度サンプリング戦略と階層的幾何学的特徴抽出モジュールを提案した。まず、3次元ガウス分布に基づく点群信頼度サンプリング戦略を設計し、奥行き推定に大きな誤差がある点に小さな信頼度を割り当て、信頼度に応じてフィルタリングする。次に、局所近傍特徴量を集約した階層的幾何学的特徴抽出モジュールと、点群中の大域的な幾何特徴量を捉えるための二重変換器を紹介する。最後に、我々の検出フレームワークは、高品質なPseudo-LiDARとエンリッチされた幾何学的特徴を入力とするPoint-VoxelRCNN(PV-RCNN)に基づいている。実験結果より、本手法は単眼3次元物体検出において満足のいく結果を得ることができた。

1. Introduction

近年、モノのインターネット[1, 2]、自動車のインターネット[3, 4]、ビッグデータとディープラーニング[5, 6]などの技術の急速な発展に伴い、都市交通も近代化とインテリジェンスに移行している。自律走行は、インテリジェント交通において最も有望なタスクの1つである[7]。第6世代(6G)セルラーネットワークの応用に成功し、自律走行は6Gの中核サービスの1つとなっている[8]。6Gは自律走行に低遅延な通信機能を提供し、車両と道路間の相乗効果を高め、自律走行の急速な発展を促進することができる。自律走行の中心的な課題は環境認識である。2次元物体検出の自動車の環境認識への応用が成功し[9]、3次元物体検出は学者の間でますます注目されている。3次元物体検出は、物体のクラス、ポーズ、正確な位置情報を得るために、物体の

多くの分野で広く使われている。fields.

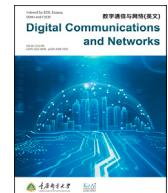
既存の3D物体検出手法は、主にLiDAR、カメラ[10]、マルチセンサーフュージョンスキームを用いて3Dデータを収集する。LiDARによる3次元物体検出は、LiDARが提供する正確な3次元点群により、最高の性能を発揮する。しかし、生成されるLiDAR点群の品質はレーザースキャンに大きく依存し、材料とプロセスによって制限されるため、コストが非常に高くなる。低コストの3D物体検出ソリューションの探索において、人々はLiDARの代わりにカメラを使い始めてコストを削減している。単眼3D(M3D)物体検出は、LiDARベースの検出方法と比較して有望であるが、奥行き情報なしにはより困難である。多くの関連研究[11, 12]が大きな進歩を遂げたが、単眼3D物体検出ではまだ解決すべき問題がある。

従来手法では、画像空間における2次元検出手法を適用することで、3次元空間における3次元パラメータを得ることができる。MonoGRNet[13]は、2次元検出の結果に従って、1枚の画像中の3次元バウンディングボックスのパラメータを予測する。

* Corresponding author.

** Corresponding author.

E-mail addresses: jlzheng@mail.xidian.edu.cn (J. Zhang), gzfang@stu.xidian.edu.cn (G. Fang), bwang@xidian.edu.cn (B. Wang), xiaobo.zhou@tju.edu.cn (X. Zhou), qqpei@mail.xidian.edu.cn (Q. Pei), cc2000@mail.xidian.edu.cn (C. Chen).



Monocular 3D object detection with Pseudo-LiDAR confidence sampling and hierarchical geometric feature extraction in 6G network

Jianlong Zhang^a, Guangzu Fang^a, Bin Wang^{a, **}, Xiaobo Zhou^b, Qingqi Pei^c, Chen Chen^{c,*}

^a School of Electronic Engineering, Xidian University, Xi'an, 710071, China

^b School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, 300000, China

^c School of Telecommunications Engineering, Xidian University, Xi'an, 710071, China

ARTICLE INFO

Keywords:

Monocular 3D object detection
Pseudo-LiDAR
Confidence sampling
Hierarchical geometric feature extraction

ABSTRACT

The high bandwidth and low latency of 6G network technology enable the successful application of monocular 3D object detection on vehicle platforms. Monocular 3D-object-detection-based Pseudo-LiDAR is a low-cost, low-power solution compared to LiDAR solutions in the field of autonomous driving. However, this technique has some problems, i.e., (1) the poor quality of generated Pseudo-LiDAR point clouds resulting from the nonlinear error distribution of monocular depth estimation and (2) the weak representation capability of point cloud features due to the neglected global geometric structure features of point clouds existing in LiDAR-based 3D detection networks. Therefore, we proposed a Pseudo-LiDAR confidence sampling strategy and a hierarchical geometric feature extraction module for monocular 3D object detection. We first designed a point cloud confidence sampling strategy based on a 3D Gaussian distribution to assign small confidence to the points with great error in depth estimation and filter them out according to the confidence. Then, we present a hierarchical geometric feature extraction module by aggregating the local neighborhood features and a dual transformer to capture the global geometric features in the point cloud. Finally, our detection framework is based on Point-Voxel-RCNN (PV-RCNN) with high-quality Pseudo-LiDAR and enriched geometric features as input. From the experimental results, our method achieves satisfactory results in monocular 3D object detection.

1. Introduction

Recently, with the rapid development of technology, e.g., the Internet of Things [1,2], the Internet of Vehicles [3,4], big data and deep learning [5,6], urban transportation has also moved toward modernization and intelligence. Autonomous driving is one of the most promising tasks in intelligent transportation [7]. With the successful application of 6th-Generation (6G) cellular networks, autonomous driving has become one of the core services of 6G [8]. 6G can provide low-latency communication capability for autonomous driving, increase the synergy between vehicles and roads, and promote the rapid development of autonomous driving. The core challenge of autonomous driving is environmental perception. With the successful application of 2D object detection to the environmental perception of vehicles [9], 3D object detection has received increasing attention from scholars. 3D object detection can obtain the class, pose and precise position information of an object, which makes it

widely used in many fields.

Existing 3D object detection methods mainly use LiDAR, cameras [10] and multisensor fusion schemes to collect 3D data. LiDAR-based 3D object detection has the best performance due to the accurate 3D point cloud provided by LiDAR. However, the quality of the generated LiDAR point cloud depends heavily on the laser scan and is limited by the material and process, which makes the cost very high. In the search for a lower-cost 3D object detection solutions, people are starting to use cameras instead of LiDAR to cut costs. Monocular 3D (M3D) object detection is promising compared to LiDAR-based detection methods, but it is more challenging without depth information. Although many associated studies [11,12] have made great progress, there are still problems to be solved in monocular 3D object detection.

Previous methods obtain 3D parameters in 3D space by applying 2D detection methods in image space. MonoGRNet [13] predicts the parameters of 3D bounding boxes in a single image according to the results

* Corresponding author.

** Corresponding author.

E-mail addresses: jlzheng@mail.xidian.edu.cn (J. Zhang), gfang@stu.xidian.edu.cn (G. Fang), bwang@xidian.edu.cn (B. Wang), xiaobo.zhou@tju.edu.cn (X. Zhou), qpei@mail.xidian.edu.cn (Q. Pei), cc2000@mail.xidian.edu.cn (C. Chen).

M3D-Region Proposal Network (RPN) [14]は、2Dと3Dの両方のボックスを最適化するためにPRN構造を利用する。これらの方法は、有効な事前知識や画像外観の合理的な制約を考慮しようとするが、3次元空間情報がないため、より良い性能を達成することができない。一方、単眼Pseudo-LiDAR法は、視覚点群をVelodyne LiDAR点群に変換することで、LiDARベースの検出ネットワークをフルに活用し、優れた検出性能を実現する。このように、最近では、単眼画像で3次元パラメータを学習する代わりに、単眼奥行き推定ネットワークによって点群を生成し、3次元点群上の3次元バウンディングボックスを直接予測するアプローチもある。Pseudo-LiDAR[15]は、3D空間におけるFrustum PointNet 3D検出フレームワーク[16]を用いて、オブジェクトの3Dバウンディングボックスを予測する。しかし、単眼深度マップ変換によって得られる擬似LiDARは、密度や深度の点でLiDARとは異なる。単眼奥行き推定は、通常、物体車両の端の背景の影響を及ぼし、推定誤差が大きい[17]。一方、誤差はシーン距離とともに大きくなる。図1より、PV-RCNNネットワークは生のPseudo-LiDARの長距離物体を検出できないことがわかる。これは、既存の特徴抽出構造によっても制限される。長距離オブジェクトに属する点の数は非常に少ないため、既存の特徴抽出構造では点群の幾何学的特徴を効果的に表現することが困難である。

ここでは、前述の問題点を解決するために、信頼度サンプリング戦略と階層的幾何学的特徴抽出モジュールを用いた単眼擬似LiDAR 3D物体検出を提案する。我々のアプローチの主なアイデアは、Pseudo-LiDARの品質を向上させ、点群の幾何学的特徴表現を改善することである。まず、Pseudo-LiDARの信頼度分布を3次元ガウス関数でフィッティングし、物体車両の端や長距離の点に小さな信頼度を割り当て、信頼度サンプリングに従って、より大きな奥行き推定誤差を持つ点(すなわち、信頼度が小さい点)をフィルタリングするPseudo-LiDAR信頼度サンプリング戦略を提案する。次に、LiDARベースの3D検出ネットワークPV-RCNN[18]を3D検出フレームワークの3D検出器として紹介する。最後に、局所的な注意特徴エンコーディングを用いて点群の局所近傍を構築し、注意重みによって特徴を集約する階層的な幾何学的特徴抽出モジュールを設計する。スカラーとベクトルの注意で作られた二重変換器を用いて、点間の大域的な相関を捉える。このように、本モジュールはPseudo-LiDARの局所的な幾何学的近傍特微量と大域的な幾何学的特微量を抽出することができる。PV-RCNN[18]のPointNetベースの集合抽象化手法と比較して、我々の特徴抽出モジュールはPseudo-LiDARの幾何学的特徴の抽出においてより効果的である。

本論文の貢献は以下のようにまとめられる。

- 3D検出フレームワークにPV-RCNNネットワークを採用することで、単眼3D物体検出のベンチマークを改善する。
- 3次元ガウス分布に基づく信頼度サンプリング戦略は、単眼奥行き推定誤差による影響を大幅に低減するために設計されている。
- 我々のモデルは、設計された階層的幾何学的特徴抽出モジュールによって抽出された点群の幾何学的特徴を得ることができる。

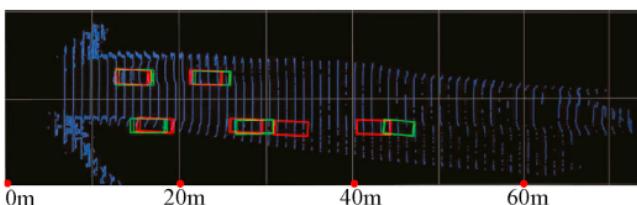


図1. PV-RCNN[18]ネットワークは、生のPseudo-LiDAR上でKITTI[19]の3D検出結果を得る。緑色のボックス: グランドトゥルース。赤枠: 予測値。

本稿の残りの部分は以下のように構成されている。セクション2では、3Dオブジェクト検出に関する関連研究を紹介する。第3節では、本手法の原理を示す。関連する実験についてはセクション4でレビューする。セクション5で我々の研究を要約する。

2. 関連研究

2.1. LiDARによる3次元物体検出

我々のインスピレーションは、LiDARベースの3Dオブジェクト検出から得られる。LiDAR点群は、RGB色情報を持つ画像データと比較してシンプルで正確であり、物体定位や姿勢推定に多くの利点を持つ。したがって、点群データの処理は、3次元物体検出アルゴリズムのカーネルタスクとなる。これらの方法は、点ベースとボクセルベースに分類することができる。最初の方法は、点群フォーマットを用いて3Dシーンを表現するもので、この方法の主な研究は以下の通りである。Frustum PointNet [16]は、2D検出ボックスを3Dフラストラム提案に投影し、PointNet [20]を用いてフラストラム提案の前景と背景を分割し、背景点の干渉を減衰させる。PointRCNN[21]は、PointNet[22]を用いてシーンの前景と背景をセグメンテーションし、高品質な3D提案を生成し、点群のプールされた特徴に基づいて提案を最適化する。

もう一つの方法は、点群をボクセルに変換し、ボクセルから特徴を抽出して3次元物体検出を行う方法である。例えば、SECOND [23]は、点群のスペース性を完全に利用する空間的に疊み込みネットワークを用いることで、点群特徴抽出能力を向上させ、学習速度を加速させる。高速Point R-CNN [24]は、Voxel-RPN層を通して少数のプロポーザルを生成し、各プロポーザルに対してよりきめ細かい回帰を実行する。PV-RCNN[18]は、ボクセルベースのアプローチにより高品質なプロポーザルを生成し、ポイントベースのアプローチにより点群の局所特徴を豊かにし、生成されたプロポーザルをさらに洗練させる。LiDARベースの3Dオブジェクト検出は有望な検出性能を達成するが、高品質のLiDAR点群を収集することは非常に高価である。

2.2. 単眼3D物体検出

LiDARベースの物体検出手法とは異なり、単眼3D物体検出は、3D空間情報を持たない單一画像を利用するために、3D検出タスクにおける重要な課題である。以前の研究では、3Dバウンディングボックスを予測するために成熟した2D検出器が採用され、その後、バウンディングボックスのパラメータが幾何学的制約で最適化された。Mono3D[25]は、3D空間における3D提案を抽出するために事前知識を利用し、2D画像空間に投影する。次に、投影されたボックス内の特微量をSVM分類に利用し、さらに提案の微調整を行う。Deep MANTA [26]は、車両特徴領域内のキーポイントを検出することで、3Dバウンディングボックスの情報を取得する。RoI-10D[27]は、微分可能な2D-RoIリフト構造を導入し、2D-RoIによってRGB画像と単眼深度特徴を融合し、3Dバウンディングボックスを得る。幾何学的関係制約や奥行き情報を用いることで、3Dオブジェクト検出性能が若干向上したものの、これらの手法は対応する3D空間情報の欠如により、まだ限界がある。

単眼奥行き推定ネットワークの急速な発展により、豊富な単眼奥行き情報を得ることができるようになった。そこで、単眼深度マップを擬似LiDAR点群に変換し、3次元空間情報を表現する。AM3D[28]は、2次元検出ネットワークを用いて深度マップ上の物体車両領域を捕捉し、2次元画像から色情報を持つ点群に物体車両を投影し、PointNetを用いて3次元バウンディングボックスを回帰する。Pseudo-LiDAR [15]は、深度推定ネットワークから得られた深度マップを直接Pseudo-LiDARに変換する。LiDARベースの3D物体検出手法は、Frustum PointNet [16]などの擬似LiDARに適用できる。

of 2D detection. M3D-Region Proposal Network (RPN) [14] utilizes PRN structures to optimize both 2D and 3D boxes. These methods cannot achieve better performance due to the lack of 3D spatial information, although they attempt to take some valid prior knowledge or reasonable constraints of image appearance. In contrast, the monocular Pseudo-LiDAR method makes full use of the LiDAR-based detection network by converting the visual point cloud to a Velodyne LiDAR point cloud and achieves excellent detection performance. Thus, recently, some approaches generated point clouds through monocular depth estimation networks and directly predicted 3D bounding boxes on 3D point clouds instead of learning 3D parameters in monocular images. Pseudo-LiDAR [15] uses the Frustum PointNet 3D detection framework [16] in 3D space to predict 3D bounding boxes of objects.

However, the Pseudo-LiDAR obtained by monocular depth map transformation is different from the LiDAR in terms of density and depth. The monocular depth estimation is usually affected by the background at the edge of the object vehicle with a large estimation error [17]; meanwhile, the error increases with the scene distance. From Fig. 1, we can observe that the PV-RCNN network fails to detect the long-distance objects in the raw Pseudo-LiDAR. This is also limited by the existing feature extraction structure. The number of points belonging to long-distance objects is very low, so the existing feature extraction structure has difficulty effectively representing the geometric features of the point cloud.

Herein, monocular Pseudo-LiDAR 3D object detection with a confidence sampling strategy and a hierarchical geometric feature extraction module is proposed to address the aforementioned problems. The main idea of our approach is to improve the quality of Pseudo-LiDAR and improve the geometric feature representation of point clouds. First, we propose a Pseudo-LiDAR confidence sampling strategy that fits the confidence distribution of Pseudo-LiDAR by a 3D Gaussian function to assign small confidence to points at the edges of object vehicles and at long distances and filter out points with larger depth estimation errors (i.e., points with small confidence) according to the confidence sampling. Then, we introduce a LiDAR-based 3D detection network PV-RCNN [18] as the 3D detector of our 3D detection framework. Finally, we design a hierarchical geometric feature extraction module that uses local attention feature encoding to construct the local neighborhood of the point cloud and aggregates the features by attention weights. It uses a dual transformer built with scalar and vector attention to capture the global correlation between points. Thus, the module can extract local geometric neighborhood features and global geometric features of Pseudo-LiDAR. Compared with the PointNet-based set abstraction method in PV-RCNN [18], our feature extraction module is more effective in extracting the geometric features of Pseudo-LiDAR.

The contributions of this paper can be summarized as follows.

- The benchmark of monocular 3D object detection is improved by employing the PV-RCNN network for the 3D detection framework.
- A confidence sampling strategy based on a 3D Gaussian distribution is designed to greatly reduce the impact caused by monocular depth estimation errors.
- Our model can obtain the geometric features of the point cloud extracted by the designed hierarchical geometric feature extraction module.

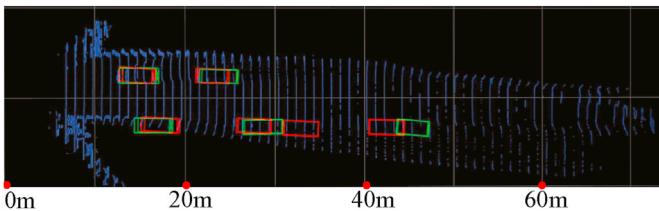


Fig. 1. PV-RCNN [18] network obtains KITTI [19] 3D detection results on raw Pseudo-LiDAR. Green boxes: ground truth. Red boxes: predicted.

The remainder of the paper is organized as follows. In Section 2, we present related work on 3D object detection. Section 3 shows the principles of our method. The relevant experiments are reviewed in Section 4. We summarize our work in Section 5.

2. Related work

2.1. LiDAR-based 3D object detection

Our inspiration comes from LiDAR-based 3D object detection. The LiDAR point cloud is simple and accurate compared with image data with RGB color information and possesses more advantages for object localization and pose estimation. Therefore, processing the point cloud data becomes the kernel task of the 3D object detection algorithm. These methods can be categorized into point-based and voxel-based methods. The first method uses the point cloud format to represent 3D scenes, and the main studies of this method are as follows. Frustum PointNet [16] projects the 2D detection box into a 3D frustum proposal and then uses PointNet [20] to segment the foreground and background of the frustum proposal to attenuate the interference of background points. PointRCNN [21] segments the foreground and background of the scene using PointNet++ [22] to generate a high-quality 3D proposal and then optimizes the proposal based on the pooled features of the point cloud.

The other method converts the point cloud into voxels to extract features from the voxels for 3D object detection. For example, SECOND [23] improves the point cloud feature extraction ability and accelerates the speed of the training by using a spatially sparse convolutional network that fully exploits the sparsity of the point cloud. The fast Point R-CNN [24] generates a small number of proposals through the Voxel-RPN layer and performs a more fine-grained regression on each proposal. PV-RCNN [18] generates high-quality proposals by the voxel-based approach, enriches the local features of the point cloud by the point-based approach, and further refines the generated proposals. While LiDAR-based 3D object detection achieves promising detection performance, collecting high-quality LiDAR point clouds is very expensive.

2.2. Monocular 3D object detection

Unlike LiDAR-based object detection methods, monocular 3D object detection utilizes a single image without 3D spatial information, which is a significant challenge in 3D detection tasks. In earlier works, mature 2D detectors were adopted to predict the 3D bounding box, and then the parameters of the bounding box were optimized with the geometric constraint. Mono3D [25] utilizes prior knowledge to extract 3D proposals in 3D space and projects them into 2D image space. Then, it uses the features within the projected box for SVM classification and further fine-tunes the proposal. Deep MANTA [26] obtains information of the 3D bounding box by detecting the key points in the vehicle feature areas. RoI-10D [27] introduces a differentiable 2D-RoI lift structure and fuses RGB images and monocular depth features by 2D-RoI to obtain 3D bounding boxes. Despite some improvement in 3D object detection performance achieved by using geometric relationship constraints or depth information, these methods are still limited by the lack of corresponding 3D spatial information.

Benefiting from the rapid development of the monocular depth estimation network, one can obtain abundant monocular depth information. Therefore, we transform the monocular depth map into a Pseudo-LiDAR point cloud to represent the 3D spatial information. AM3D [28] captures the object vehicle regions on the depth map by using a 2D detection network, projects the object vehicles from the 2D image into the point cloud with color information, and regresses the 3D bounding boxes with PointNet. Pseudo-LiDAR [15] directly transforms the obtained depth map from the depth estimation network into Pseudo-LiDAR. LiDAR-based 3D object detection methods can be applied to Pseudo-LiDAR, e.g., Frustum PointNet [16]. RefinedMPL [29] uses a supervised and an

RefinedMPL[29]は、教師あり、教師なしの手法で前景点を区別し、干渉する背景点をフィルタリングすることで、Pseudo-LiDAR点群の高密度化と画像エッジ近傍の背景点が検出を妨害する状況に対処する。関連する実験では、単眼画像に対する優れた検出性能が示された。

単眼擬似LiDAR 3D物体検出は、単眼画像における2次元特徴抽出問題を、単眼奥行き推定と点群特徴抽出の問題に変換する。成熟した単眼奥行き推定ネットワークと高性能なLiDARベースの3Dオブジェクト検出ネットワークを使用することで、より優れた単眼3D検出性能が得られると考えられる。RefinedMPL[29]において、背景点が検出性能に影響することが検証されている。そこで、Pseudo-LiDARの生成品質を向上させ、既存のLiDARベースの3D物体検出ネットワークを最適化するために、新しい単眼3D物体検出法を提案した。

3. Approach

単眼画像から3次元バウンディングボックスを得るために単眼Pseudo-LiDAR 3D検出フレームワークを提案した。既存の単眼擬似LiDARのための3D検出手法、例えばRefinedMPL [29]は、主に多数の背景点をフィルタリングし、高性能なLiDARベースの3D検出ネットワークを利用することで、検出性能を向上させている。ほとんどの場合、高品質な点群は3D検出器の計算コストを削減し、高品質な提案を生成することができるため、既存のLiDARベースの3D検出ネットワークは、単眼擬似LiDAR 3D物体検出において優れた検出性能を達成できる。

提案手法は、擬似LiDAR信頼度サンプリング戦略と階層的幾何学的特徴抽出モジュールを適用し、生成された擬似LiDAR点群の品質を向上させ、LiDARベースの検出ネットワークによる検出性能を向上させる。図2に示すように、我々のアプローチには3つの段階がある。Pseudo-LiDAR生成段階では、DORN[30]を用いて深度情報を取得する。得られた奥行き推定マップから、座標変換と投影変換を用いてPseudo-LiDARを得る。Pseudo-LiDARの信頼度サンプリングでは、信頼度サンプリングを用いて高品質のPseudo-LiDAR点群を生成する。まず、シーン分布に従って、Pseudo-LiDARの局所的信頼度分布と大域的信頼度分布をフィッティングする。次に、単眼奥行き推定誤差が大きい点を信頼度サンプリングによりフィルタリングし、より高品質な擬似LiDAR点群を得る。第3段階、すなわち擬似LiDAR物体検出では、満足のいく検出性能を得るために、階層的な幾何学的特徴抽出モジュールを設計する。

この段階では、PV-RCNNを基幹ネットワークとして使用する。次に、PV-RCNN[18]におけるキーポイントの幾何学的特徴表現能力が弱いという問題に対処するため、局所的注意特徴符号化と二重変換器を用いて、それぞれPseudo-LiDARの局所的幾何学的特徴と大域的幾何学的特徴を取得する。最後に、最終的な検出結果はPV-RCNNのバウンディングボックス予測ネットワークによって得られる。

3.1. 擬似LiDAR生成

3.1.1. 単眼奥行き推定

単眼奥行き推定ネットワークは、画像を入力として、ピクセル単位の奥行き情報を出力する。深度推定の精度はPseudo-LiDAR点群の品質を決定するため、深度マップDを得るために、成功したDORN [30]を単眼深度推定ネットワークとして採用する。DORNアルゴリズムのフレームワークは、点群の深度点推定の精度が低いという問題を解決するために順序回帰を導入する[30]。

3.1.2. 擬似LiDAR生成

深度マップDは、RGB画像空間の深度情報のみを取得する。Pseudo-LiDAR点群を生成するためには、座標変換と投影変換によって奥行き情報をさらに取得する必要がある。そのプロセスを図2に示す。深度マップは離散ピクセル座標系で、Pseudo-LiDARは連続物理座標系である。擬似LiDAR点群の3次元座標(x, y, z)は、各画素座標(u, v)に投影変換[15]を施すことにより、以下の式により求められる：

$$\begin{cases} x = \frac{z}{f_c}x', x' = s_x(u - u_0) \\ y = \frac{z}{f_c}y', y' = s_y(v - v_0) \\ z = D(x', y') \end{cases} \quad (1)$$

ここで、 x^0 と y^0 は物理的距離の座標、 s_x と s_y は任意の2画素間の物理的距離、 (u_0, v_0) は画素座標系における物理座標系の原点の位置である。 f_c はカメラの焦点距離である。

3.2. Confidence sampling

擬似LiDAR生成プロセスでは、奥行き情報が点群にエンコードされる。しかし、単眼式奥行き推定ネットワークの限界により、対象車両の端部でも、距離の長い領域でも、奥行き推定誤差が大きくなる。

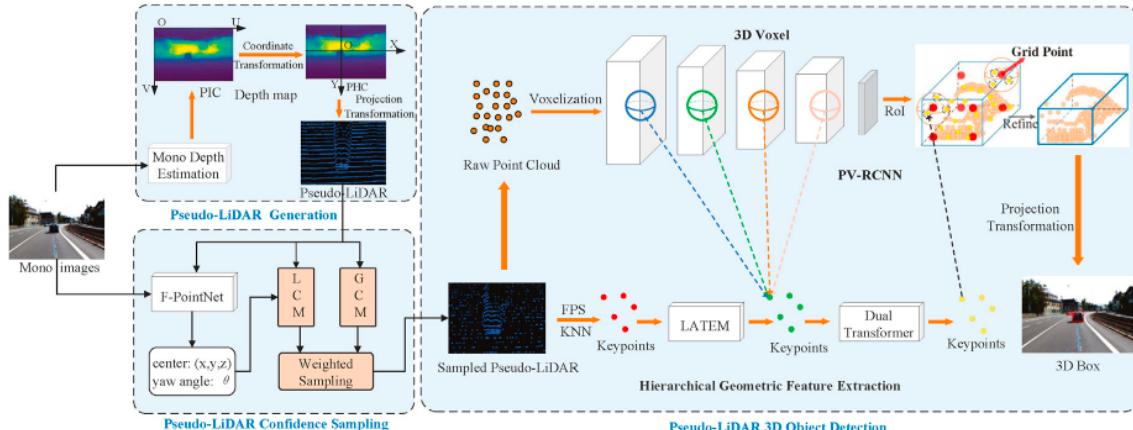


図2. 我々のフレームワークの全体アーキテクチャPIC: PIC: ピクセル座標系; PHC: 物理座標系; FPS: 最遠点サンプリング; LATEM: 局所注意特徴符号化モジュール。

unsupervised method to distinguish the foreground points and then filters out the interfering background points to address the situation in which the high density of Pseudo-LiDAR point clouds and many background points near image edges interfere with the detection. The associated experiments show excellent detection performance on monocular images.

Monocular Pseudo-LiDAR 3D object detection transforms the 2D feature extraction problem in monocular images into the problems of monocular depth estimation and point cloud feature extraction. Better monocular 3D detection performance is supposed to be obtained by using a mature monocular depth estimation network and a high-performance LiDAR-based 3D object detection network. It was verified in RefinedMPL [29] that the background points influence the detection performance. Therefore, to improve the generation quality of Pseudo-LiDAR and optimize existing LiDAR-based 3D object detection networks, we proposed a novel monocular 3D object detection method.

3. Approach

We proposed a monocular Pseudo-LiDAR 3D detection framework for obtaining 3D bounding boxes from monocular images. Existing 3D detection methods for monocular Pseudo-LiDAR, e.g., RefinedMPL [29], mainly improve the detection performance by filtering a large number of background points and utilizing a high-performance LiDAR-based 3D detection network. In most cases, the high-quality point clouds can reduce the computational cost of the 3D detector and generate high-quality proposals, which is why the existing LiDAR-based 3D detection networks can achieve excellent detection performance in monocular Pseudo-LiDAR 3D object detection.

Our proposed method applies a Pseudo-LiDAR confidence sampling strategy and a hierarchical geometric feature extraction module to improve the quality of the generated Pseudo-LiDAR point clouds and enhance the detection performance with the LiDAR-based detection network. As shown in Fig. 2, our approach has three stages. In the Pseudo-LiDAR generation stage, we employ DORN [30] to obtain depth information. We obtain Pseudo-LiDAR from the obtained depth estimation map by using coordinate transformation and projection transformation. In Pseudo-LiDAR confidence sampling, we use confidence sampling to generate high-quality Pseudo-LiDAR point clouds. First, we fit the local and global confidence distribution of the Pseudo-LiDAR according to the scene distribution. Then, the points with large monocular depth estimation errors are filtered out by confidence sampling to obtain higher quality Pseudo-LiDAR point clouds. In the third stage, i.e., Pseudo-LiDAR object detection, we design a hierarchical geometric feature extraction module to obtain satisfactory detection performance.

In this stage, we use PV-RCNN as our backbone network. Then, to address the problem of weak geometric feature representation capability for keypoints in PV-RCNN [18], we use the local attention feature encoding and the dual transformer to acquire the local and global geometric features of the Pseudo-LiDAR, respectively. Finally, the final detection results are obtained by the bounding box prediction network of the PV-RCNN.

3.1. Pseudo-LiDAR generation

3.1.1. Monocular depth estimation

The monocular depth estimation network obtains pixelwise depth information output by using the image as input. Because the accuracy of depth estimation determines the quality of the Pseudo-LiDAR point clouds, we adopt the successful DORN [30] as our monocular depth estimation network to obtain the depth map D . The DORN algorithmic framework introduces ordered regression to solve the problem of less accurate depth point estimation for point clouds [30].

3.1.2. Pseudo-LiDAR generation

The depth map D only obtains the depth information of the RGB image space. Depth information needs to be further obtained by coordinate transformation and projection transformation to generate the Pseudo-LiDAR point cloud. The process is shown in Fig. 2. The depth map is in a discrete pixel coordinate system, and the Pseudo-LiDAR is in a continuous physical coordinate system. The 3D coordinates (x, y, z) of the Pseudo-LiDAR point cloud are obtained by applying a projection transformation [15] to each pixel coordinate (u, v) according to the following equations:

$$\begin{cases} x = \frac{z}{f_c}, x' = s_x(u - u_0) \\ y = \frac{z}{f_c}, y' = s_y(v - v_0) \\ z = D(x', y') \end{cases} \quad (1)$$

where x' and y' are the coordinates of the physical distance, s_x and s_y are the physical distances between two arbitrary pixels, and (u_0, v_0) is the position of the origin of the physical coordinate system in the pixel coordinate system. f_c is the camera focal length.

3.2. Confidence sampling

In the Pseudo-LiDAR generation process, the depth information is encoded into the point cloud. However, the limitations of the monocular depth estimation network result in large depth estimation errors both at

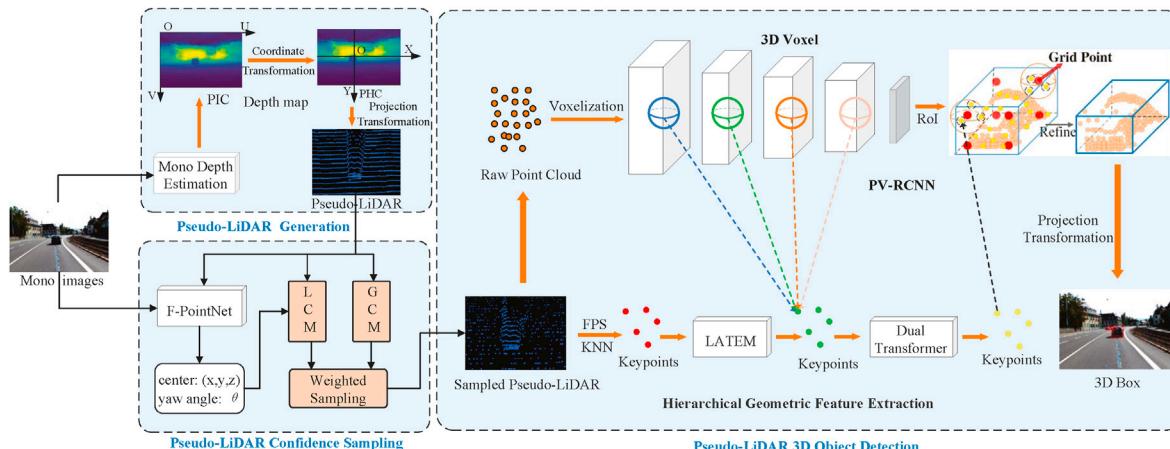


Fig. 2. Overall architecture of our framework. PIC: Pixel Coordinate System; PHC: Physical Coordinate System; FPS: Farthest Point Sampling; LATEM: Local Attention Feature Encoding Module.

奥行き推定誤差の影響を軽減するために、まず各ポイントに対する信頼度を割り当てる。物体車両の3次元中心点と点間の距離から局所信頼度を算出する。そして、シーンの距離分布に従って、グローバルな信頼度を求める。最後に、Pseudo-LiDAR信頼度は、大域的信頼度を局所的信頼度で重み付けすることで得られ、信頼度分布に従って点群がサンプリングされる。

3.2.1. 3次元中心点・方向推定

局所的な信頼性は、オブジェクト車両内の点と3D中心点との間の距離の分布を必要とする。したがって、対象車両のヨー角と3次元中心点を知ることで、対象車両の内側にあるかどうかを判断し、点間の距離を知る必要がある。F-PointNetネットワークを用いて、物体車両の3次元中心点(x_c, y_c, z_c)とヨー角 θ を推定する。F-PointNet[16]は、2次元画像中の2次元CNNによって3次元Frustum提案を生成し、軽量回帰PointNetを利用して(x_c, y_c, z_c)と θ を推定する。

3.2.2. 局所信頼度生成

物体車両の端部では単眼奥行き推定誤差が大きいため、物体車両内の各点に重みを割り当て、それを局所信頼度と呼ぶ。点が3次元中心点に近いほど奥行き推定誤差が小さくなることを考慮し、Pseudo-LiDARの局所信頼度をモデル化するために3次元ガウス関数を利用する。したがって、信頼度が大きいほど誤差は小さくなり、逆もまた然りである。KITTIデータセットの車両のサイズは類似しているので、各車両の平均長、幅、高さをとり、ある点が対象車両の内側にあるか外側にあるかを判断する。

擬似LiDAR点群における点の3次元座標を3次元空間における $p(x, y, z)$ とする。計算を容易にするため、車両の頭部方向をx軸とする。車両のヨー角は θ であるため、擬似LiDAR座標系を反時計回りに θ 度回転させ、以下のように新しい座標表現を得る：

$$\begin{cases} x' = x \cos \theta + y \sin \theta \\ y' = -x \sin \theta + y \cos \theta \\ z' = z \end{cases} \quad (2)$$

ここで、 $p(x, y, z)$ と3次元中心点(x_c, y_c, z_c)の新しい座標は、それぞれ (x^0, y^0, z^0) と $\delta x_c^0; y_c^0; z_c^0$ である。

車両 b に対する点 p の3次元ガウス重みは

$$\alpha_{(p,b)} = \begin{cases} s(x', y', z') & y(p, b) = 1, \\ 0 & y(p, b) = 0. \end{cases} \quad (3)$$

ここで、 $y(p, b) = 1/4$ は、 p が車両 b の内側にない点である。そして、 y の計算式は以下になる：

$$y(p, b) = \begin{cases} 0 & \text{if } |x'_c - x'| > l/2 \\ 0 & \text{if } |y'_c - y'| > w/2 \\ 0 & \text{if } |z'_c - z'| > h/2 \\ 1 & \text{otherwise.} \end{cases} \quad (4)$$

ここで、 $\delta x_c; y_c; z_c$ は回転変換後の推定3次元中心点である。 l, w, h はそれぞれ b の平均長さ、幅、高さである。

点 p が車両 b の内側にある場合、車両の内側にある信頼度重みを3次元ガウス関数で計算すると、以下になる：

$$s(x', y', z') = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3} e^{-\frac{(x'-x_c)^2 + [(y'-y_c)\frac{l}{w}]^2 + [(z'-z_c)\frac{h}{w}]^2}{2\sigma^2}} \quad (5)$$

ここで、 σ は3次元ガウス関数の減衰率パラメータである。そして、点 p の局所信頼度 $S_{Local}(p)$ の式は

$$S_{Local}(p) = \max(\lambda_\alpha \cdot f_{norm}(\alpha_{(p,b)}), \xi_\alpha) \quad (6)$$

ここで、 λ_α は重みバランスパラメータである。 f_{norm} は重み正規化関数で、 p の信頼度が0から1の間に保証される。 ξ_α は背景信頼度の閾値である。

3.2.3. グローバル信頼度生成

単眼画像におけるステレオ視覚情報の欠如は、長距離シーンの単眼奥行き推定における性能の低下につながり、長距離シーンでは奥行き推定誤差が増大する。この問題を解決するために、長距離点の信頼度を信頼度減衰率 R_γ で減少させる。自律走行シーンの多様性により、奥行き分布は大きく変化する。そこで、異なるシーンに対する R_γ を以下のように計算する。

$$R_\gamma = \frac{1}{\lambda_\beta f_E(Q) + f_D(Q)} \quad (7)$$

ここで、 Q は現在のシーンの点群集合である。 λ_β は Q の平均と分散の重みを調整するためのグローバルバランスパラメータであり、 $f_E(Q)$ と $f_D(Q)$ はそれぞれ Q の平均と分散を表す。

$$f_E(Q) = \frac{\sum_{p \in Q} d_p}{|Q|}, f_D(Q) = \sqrt{\frac{\sum_{p \in Q} [d_p - f_E(Q)]^2}{|Q|}} \quad (8)$$

ここで、 d_p は p の深度値である。(7)は、 Q の平均と分散が大きく、信頼度が減衰しそうないように減衰率を小さく設定することを示している。この設定により、シーン内のオブジェクトポイントが完全にフィルタリングされることを避けることができる。

信頼度減衰率 R_γ はシーン距離が大きくなるにつれて信頼度を低下させることができ、点 p の大域的信頼度 $S_{Global}(p)$ は

$$S_{Global}(p) = \max(1 - R_\gamma d_p, \xi_\beta) \quad (9)$$

ξ_β は、長距離の背景点が完全にフィルタリングされず、長距離の物体車両点がサンプリングされることを保証するためのグローバル信頼度背景閾値である。

点 p の最終的な信頼度 $S(p)$ は、局所的な信頼度よりも大域的な信頼度を重み付けすることによって生成される、

$$S(p) = S_{Local}(p) \cdot S_{Global}(p) \quad (10)$$

3.2.4. Confidence sampling

擬似LiDAR信頼度分布が得られたら、生の擬似LiDAR点群 Q_{raw} を以下のようにサンプリングしてサンプリング点群集合 Q_{sam} を求める：

$$Q_{sam} = \{p | S(p) > rand(0, 1), p \in Q_{raw}\} \quad (11)$$

$rand(0, 1)$ は0と1の間の確率変数を生成できる関数である。この式は、点 p の信頼度が $rand(0, 1)$ より大きいとき、点 p は保持され、逆に p はフィルタリングされることを示している。背景点のような信頼度の低い点はフィルタリングされていることがわかる。

3.3. 階層的な幾何学的特徴抽出を用いた擬似LiDAR 3次元物体検出ネットワーク

本節では、PV-RCNNネットワークを検出ネットワークの3次元検出器として取り上げる。PV-RCNN[18]では、シーン全体の全てのボクセル特徴を少数のキーポイントにエンコードすることが最も重要である。しかし、この結果、キーポイントの幾何学的特徴が鈍感になる。PV-RCNNにおけるキーポイントの特徴情報をさらに充実させるために、階層的な幾何学的特徴抽出モジュールを設計した。

the edges of the object vehicle and in regions with long distances. To reduce the influence of the depth estimation error, we first assign the corresponding confidence to each point. We calculate the local confidence based on the distance between the points and the 3D center point of the object vehicle. Then, the global confidence is obtained according to the distance distribution of the scene. Finally, Pseudo-LiDAR confidence is obtained by weighting the global confidence with the local confidence, and the point cloud is sampled according to the confidence distribution.

3.2.1. 3D center point and direction estimation

The local confidence requires the distribution of the distance between the points and the 3D center point inside the object vehicle. Therefore, we need to know the yaw angle and 3D center point of the object vehicle to determine whether the point is inside the object vehicle and to know the distance between points. We use the F-PointNet network to estimate the 3D center point (x_c, y_c, z_c) and yaw angle θ of the object vehicle. F-PointNet [16] generates a 3D Frustum proposal by a 2D CNN in a 2D image and then utilizes a lightweight regression PointNet to estimate (x_c, y_c, z_c) and θ .

3.2.2. Local confidence generation

Due to the large monocular depth estimation error at the edges of the object vehicle, we assign a weight to each point within the object vehicle and call it the local confidence. Considering that the closer the point to the 3D center point, the smaller its depth estimation error, we utilize a 3D Gaussian function to model the local confidence of Pseudo-LiDAR. Therefore, the larger its confidence, the smaller its error, and vice versa. Since the sizes of the vehicles in the KITTI dataset are similar, we take the average length, width and height of each vehicle to determine whether a point is inside or outside the object vehicles.

The 3D coordinates of a point in a Pseudo-LiDAR point cloud are denoted by $p(x, y, z)$ in 3D space. To facilitate the calculation, we use the vehicle's head direction as the x-axis. Since the vehicle has a yaw angle of θ , we rotate the Pseudo-LiDAR coordinate system by θ degrees in a counterclockwise direction to obtain the new coordinate representation as follows:

$$\begin{cases} x' = x \cos \theta + y \sin \theta \\ y' = -x \sin \theta + y \cos \theta \\ z' = z \end{cases} \quad (2)$$

where the new coordinates of $p(x, y, z)$ and 3D center point (x_c, y_c, z_c) are (x', y', z') and (x'_c, y'_c, z'_c) , respectively.

The 3D Gaussian weight of point p with respect to vehicle b is

$$\alpha_{(p,b)} = \begin{cases} s(x', y', z') & y(p, b) = 1, \\ 0 & y(p, b) = 0. \end{cases} \quad (3)$$

Here, $y(p, b) = 0$ is the point where p is not inside vehicle b . Then, the expression for the calculation of y is as follows:

$$y(p, b) = \begin{cases} 0 & \text{if } |x'_c - x'| > l/2 \\ 0 & \text{if } |y'_c - y'| > w/2 \\ 0 & \text{if } |z'_c - z'| > h/2 \\ 1 & \text{otherwise.} \end{cases} \quad (4)$$

Here, (x'_c, y'_c, z'_c) is the estimated 3D center point after the rotation transformation. l, w and h are the average length, width and height of b , respectively.

If point p is inside vehicle b , we calculate the confidence weights inside the vehicle by a 3D Gaussian function as follows:

$$s(x', y', z') = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3} e^{-\frac{(x'-x'_c)^2 + [(y'-y'_c)\frac{l}{h}]^2 + [(z'-z'_c)\frac{w}{h}]^2}{2\sigma^2}} \quad (5)$$

where σ is the decay rate parameter of the 3D Gaussian function.

Then, the expression of the local confidence $S_{Local}(p)$ of point p is

$$S_{Local}(p) = \max(\lambda_a \cdot f_{norm}(\alpha_{(p,b)}), \xi_a) \quad (6)$$

Here, λ_a is the weight balance parameter. f_{norm} is the weight normalized function, which ensures that the confidence of p is between 0 and 1. ξ_a is the background confidence threshold.

3.2.3. Global confidence generation

Lack of stereo visual information in monocular images leads to poor performance in monocular depth estimation of long-distance scenes, and the depth estimation error increases in the long-distance scenes. To solve this problem, we decrease the confidence of long-distance points with the confidence decay rate R_γ . Due to the diversity of autonomous driving scenes, the depth distribution varies greatly. Therefore, we calculate R_γ for different scenes as follows:

$$R_\gamma = \frac{1}{\lambda_\beta f_E(Q) + f_D(Q)} \quad (7)$$

Here, Q is a point cloud set of the current scene. λ_β is a global balance parameter to adjust the weight of the mean and variance of Q . $f_E(Q)$ and $f_D(Q)$ denote the mean and variance of Q , respectively.

$$f_E(Q) = \frac{\sum_{p \in Q} d_p}{|Q|}, f_D(Q) = \sqrt{\frac{\sum_{p \in Q} [d_p - f_E(Q)]^2}{|Q|}} \quad (8)$$

Here, d_p is the depth value of p . (7) shows that the mean and variance of Q are large, and a smaller decay rate is set to prevent the confidence from decaying too much. This setting avoids the object points in the scene being completely filtered out.

The confidence decay rate R_γ can make the confidence decrease as the scene distance increases, and the global confidence $S_{Global}(p)$ of point p is

$$S_{Global}(p) = \max(1 - R_\gamma d_p, \xi_\beta) \quad (9)$$

ξ_β is the global confidence background threshold to ensure that long-distance background points are not completely filtered out and long-distance object vehicle points are sampled.

The final confidence $S(p)$ of point p is generated by weighting the global confidence over the local confidence, i.e.,

$$S(p) = S_{Local}(p) \cdot S_{Global}(p) \quad (10)$$

3.2.4. Confidence sampling

Once the Pseudo-LiDAR confidence distribution is obtained, the sampled point cloud set Q_{sam} is obtained by sampling the raw Pseudo-LiDAR point cloud Q_{raw} as follows:

$$Q_{sam} = \{p | S(p) > rand(0, 1), p \in Q_{raw}\} \quad (11)$$

$rand(0, 1)$ is a function that can generate a random variable between 0 and 1. This equation shows that when the confidence of point p is greater than $rand(0, 1)$, then point p is retained, and vice versa, p is filtered out. It can be seen that points with small confidence, such as background points, are filtered out.

3.3. Pseudo-LiDAR 3D object detection network with hierarchical geometric feature extraction

In this section, we take the PV-RCNN network as a 3D detector for our detection network. In PV-RCNN [18], its most critical point is to encode all voxel features of the entire scene into a small number of keypoints. However, this results in insensitivity to the geometric features of keypoints. To further enrich the feature information of keypoints in PV-RCNN, we designed a hierarchical geometric feature extraction

生の点群からキーポイントを抽出した後、Local Attention Feature Encoding (LAFFE)を用いて各キーポイントの局所幾何学的近傍を構築し、幾何学的近傍特徴をキーポイントに集約する。また、キーポイントの特徴情報を豊かにするために、各キーポイントに対応するRGB情報を集約する。最後に、キーポイント間のグローバルな幾何学的特徴情報を捕捉するための二重変換器を設計する。

3.3.1. 局所注意特徴符号化

特徴量の符号化のプロセスは以下の通りである。まず、点群情報を豊かにするために、各画素の色情報を割り当てる。空間情報と色情報を組み合わせる。図3に示すように、各キーポイントの空間幾何情報 f_{xyz} (点群のXYZ位置)と色情報 f_{rgb} (点群のRGB情報)を以下の式で符号化する：

$$f_e = \text{MLP}(f_{xyz}) \oplus \text{MLP}(f_{rgb}) \quad (12)$$

ここで、 f_e は符号化特徴量、MLPは多層パーセプトロン、接続演算で \oplus である。

次に、K-Nearest Neighbor (KNN)を用いて、生の点群からキーポイントの近傍点を見つけ、局所近傍点を構築する。ここで、KNNは生点からキーポイントまでのユークリッド距離に基づくクラスタリングを実現する。文献[31]に触発された。[31]に触発され、K個の近傍点を持つキーポイント i に対して、その局所近傍特徴 f_l^{ik} を以下のように符号化する。

$$f_l^{ik} = \text{MLP}\left(f_{xyz}^i \oplus f_{xyz}^i - f_{xyz}^{ik} \oplus f_e^i \oplus f_e^i - f_e^{ik}\right) \quad (13)$$

ここで、 k はキーポイント i の k 番目の近傍点を表す。 f_{xyz}^i f_{xyz}^{ik} と f_e^i f_e^{ik} はそれぞれ k 番目の近傍点の相対位置(キーポイント間の幾何学的構造特徴を表す)と相対特徴を表す。生点群からの位置情報と色情報を組み合わせることで、キーポイントの幾何学的特徴を強化する。

既存の研究で使用されている最大または平均ブーリングによって隣接する特徴を集約すると、キーポイント周辺の無関係な情報が保持され、重要な情報が失われる。重みを学習することで、有益な特徴を選択的に集約するアテンションメカニズムを導入した。文献[32]に触発された。[32]に触発され、我々の注意集約アプローチは以下の通りである。キーポイント i の各近傍点について、以下のように注目度を計算する：

$$\alpha_i^k = \text{softmax}\left(\text{MLP}\left(f_{xyz}^i - f_{xyz}^{ik} \oplus f_e^i - f_e^{ik}\right)\right) \quad (14)$$

ここで、softmaxは正規化関数である。

キーポイント i の最終的な局所近傍符号化特徴量 F_{agg}^i は、以下の式で計算できる：

$$F_{agg}^i = \sum_{k \in K} \alpha_i^k * f_l^{ik} \quad (15)$$

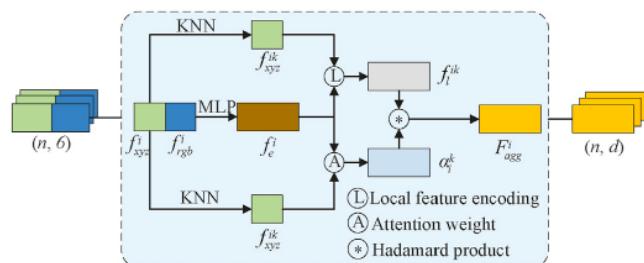


図3. 局所注意特徴符号化モジュール(MLP:多層パーセプトロン、KNN:K-最近傍)。

ここで、 K は近傍の数である。 $*$ はハダマード積である。

3.3.2. デュアル変換器による大域的な幾何学的特徴抽出

近年、変換器と自己注意は、自然言語処理[33, 34]やコンピュータビジョンの分野で大きな可能性を示している。トランスフォーマーは、シーケンスの長期的な相関を捉えることができるという利点がある。点群は3次元空間情報を表現するために使用される不規則なデータであるため、無秩序なシーケンスとみなし、変換器を適用して点間のグローバルな相関を捉える。自己注意には主に2つのタイプがある：ベクトル[35]とスカラーハadamard積[33]。スカラーハテーションは、点間の空間的な相間に注目し、文脈上の特徴の長距離表現を得ることができる。ベクトル注目の注目度はベクトルであるため、各特徴チャネルを独立に変調することができる。文献[36]に触発された。[36]に触発され、我々はこれら2種類の注意の利点を組み合わせ、キーポイントの大域的な幾何学的特徴表現を強化するための二重変換器構造を提案する。これら2つの変換器の詳細を以下に説明する。

スカラーハテーション[33]に基づく変換器の計算は、以下のように終了する。

$$\begin{cases} \mathbf{y}_{si} = \sum_{\mathbf{x}_j \in \chi} \rho(\alpha(\mathbf{x}_i)^T \psi(\mathbf{x}_j) + \delta) \beta(\mathbf{x}_j) \\ \delta = \theta(\mathbf{p}_i^T \mathbf{p}_j) \end{cases} \quad (16)$$

ここで、 χ は特徴量の集合であり、 \mathbf{y}_{si} はキーポイント i の出力特徴量である。 α 、 ψ 、 β はMLPなどの点の特徴変換である。 ρ はsoftmaxのような正規化関数である。 δ は位置符号化関数であり、より多くの文脈情報を提供するために、点対間の位置関係を取得するために使用される。 θ はMLP。 \mathbf{p}_i と \mathbf{p}_j はそれぞれ点 i と点 j の3次元座標である。

ベクトル注意とスカラーハテーションの違いは、主に注意の重みと符号化関数の計算で生じる。ベクトル注意では、[36]に従い、注意の重みと点のペア間の位置符号化関係を計算するために減算関係を採用する。位置エンコーディングは、キーポイント間の幾何学的構造をよりよく表現するために相対座標を使用する。キーポイント i の特徴出力 \mathbf{y}_{vi} は以下のように計算される：

$$\begin{cases} \mathbf{y}_{vi} = \sum_{\mathbf{x}_j \in \chi} \rho(\alpha(\mathbf{x}_i) - \psi(\mathbf{x}_j) + \delta) \beta(\mathbf{x}_j) \\ \delta = \theta(\mathbf{p}_i - \mathbf{p}_j) \end{cases} \quad (17)$$

最後に、キーポイント i の特徴は $y_f / 4$ y_{si} y_{vi} であり、2つの注意の利点を十分に利用し、キーポイントの幾何学的構造間の大域的相関を改善する。

4. Experiment

4.1. Setup

4.1.1. Dataset

本手法を一般的なKITTI[19]で検証した。データセットはトレーニング用とテスト用の画像を提供する。KITTIは現在、自律走行分野における最も重要な大規模シナリオデータセットの一つである。しかし、時間や天候に多様性がないことは、KITTIの限界である。このデータセットはテストセットのグランドトゥルースを直接提供しないため、オンラインテストには特別な要件が必要である。そこで、[12]に従い、学習セットを3712サンプル(学習セット)と3769サンプル(検証セット)の2つに分割した。

4.1.2. 評価指標

本アルゴリズムを検証するために、KITTI検証セットで3Dと鳥瞰図(BEV)の両方の物体検出結果を車について評価した。対応するPrecision/Recall曲線を計算し、閾値0.5と0.7の平均Precision(AP)を求める。KITTIは現在AP₁₁の代わりにAP₄₀を採用しているが、多くの既存手法ではAP₁₁の結果しか得られていない。

module. After the keypoints are extracted from the raw point cloud, we construct a local geometric neighborhood for each keypoint by using Local Attention Feature Encoding (LAFE) and then aggregate geometric neighborhood features to the keypoint. We also aggregate the corresponding RGB information to each keypoint to enrich the feature information of the keypoints. Finally, we design a dual transformer to capture the global geometric feature information between the keypoints.

3.3.1. Local attention feature encoding

The process of feature encoding is as follows. First, we assign the color information of each pixel to enrich the point cloud information. We combine the spatial information and color information. As shown in Fig. 3, we encode its spatial geometric information f_{xyz} (XYZ position of the point cloud) and color information f_{rgb} (RGB information of the point cloud) for each keypoint using the following equation:

$$f_e = \text{MLP}(f_{xyz}) \oplus \text{MLP}(f_{rgb}) \quad (12)$$

Here, f_e is the encoding feature, MLP is the multi-layer perceptron and \oplus is the connection operation.

Then, we use the K-Nearest Neighbor (KNN) to find the neighbor points of the keypoints in the raw point cloud to construct a local neighborhood. Herein, KNN realizes clustering based on the Euclidean distance from the raw point to the keypoint. Inspired by Ref. [31], for keypoint i with K neighbor points, we encode its local neighborhood feature f_l^{ik} as

$$f_l^{ik} = \text{MLP}\left(f_{xyz}^i \oplus f_{xyz}^i - f_{xyz}^{ik} \oplus f_e^i \oplus f_e^i - f_e^{ik}\right) \quad (13)$$

where k denotes the k -th neighbor of the keypoint i . $f_{xyz}^i - f_{xyz}^{ik}$ and $f_e^i - f_e^{ik}$ denote the relative position (representing the geometric structural features between keypoints) and relative features of the k -th neighbor point, respectively. We enhance the geometric features of keypoints by combining the position information and color information from the raw point cloud.

Aggregating neighboring features by maximum or average pooling used in existing works results in irrelevant information around keypoints being retained and important information being lost. We introduced an attention mechanism to selectively aggregate beneficial features by learning weights. Inspired by Ref. [32], our attentional aggregation approach is as follows. The attention weight is calculated for each neighbor point of keypoint i as follows:

$$\alpha_i^k = \text{softmax}\left(\text{MLP}\left(f_{xyz}^i - f_{xyz}^{ik} \oplus f_e^i - f_e^{ik}\right)\right) \quad (14)$$

where softmax is a normalization function.

The final local neighborhood encoding feature F_{agg}^i of keypoint i can be calculated by the following equation:

$$F_{agg}^i = \sum_{k \in K} \alpha_i^k * f_l^{ik} \quad (15)$$

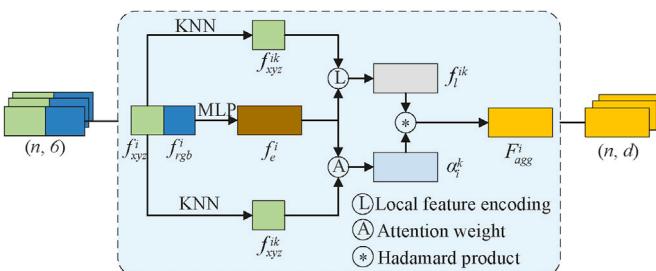


Fig. 3. Local attention feature encoding module. (MLP: Multi-Layer Perceptron, KNN: K-Nearest Neighbor.)

where K is the number of neighbors. $*$ is the Hadamard product.

3.3.2. Global geometric feature extraction with dual transformer

Recently, transformers and self-attention have shown great potential in the field of NLP [33,34] and computer vision. Transformers have the advantage of capturing long-term correlations in sequences. Since point clouds are irregular data used to represent 3D spatial information, we consider them as a disordered sequence, and then apply transformers to capture the global correlation between points. There are two main types of self-attention: vector [35] and scalar [33]. Scalar attention focuses on the spatial correlation between points and obtains the long-range representation of contextual features. The attention weight of vector attention is a vector, so it can independently modulate each feature channel. Inspired by Ref. [36], we combine the advantages of these two types of attention and propose a dual transformer structure to enhance the global geometric feature representation of keypoints. The details of these two transformers are described as follows.

The computation of the transformer based on scalar attention [33] is finished by

$$\begin{cases} \mathbf{y}_{si} = \sum_{\mathbf{x}_j \in \chi} \rho(\alpha(\mathbf{x}_i)^T \psi(\mathbf{x}_j) + \delta) \beta(\mathbf{x}_j) \\ \delta = \theta(\mathbf{p}_i^T \mathbf{p}_j) \end{cases} \quad (16)$$

Here, χ is the set of features. \mathbf{y}_{si} is the output feature of keypoint i . α , ψ and β is a feature transformation for points such as MLP. ρ is a normalization function such as softmax. δ is a position encoding function that is used to acquire the position relationship between point pairs to provide more contextual information. θ is an MLP, and \mathbf{p}_i and \mathbf{p}_j are the 3D coordinates of points i and j , respectively.

The differences between vector attention and scalar attention occur mainly in the calculation of attention weights and encoding functions. In vector attention, we adopt subtractive relations to calculate the attentional weights and positional encoding relations between pairs of points following [36]. The position encoding uses relative coordinates to better represent the geometric structure between the keypoints. The feature output \mathbf{y}_{vi} of keypoint i is calculated as follows:

$$\begin{cases} \mathbf{y}_{vi} = \sum_{\mathbf{x}_j \in \chi} \rho(\alpha(\mathbf{x}_i) - \psi(\mathbf{x}_j) + \delta) \beta(\mathbf{x}_j) \\ \delta = \theta(\mathbf{p}_i - \mathbf{p}_j) \end{cases} \quad (17)$$

Finally, the feature of keypoint i is $\mathbf{y}_f = \mathbf{y}_{si} \oplus \mathbf{y}_{vi}$, which fully exploits the advantages of the two attentions and improves the global correlation between the geometric structures of keypoints.

4. Experiment

4.1. Setup

4.1.1. Dataset

We verified our method on the popular KITTI [19]. The dataset provides images for training and testing. KITTI is currently one of the most important large-scale scenario datasets in the field of autonomous driving. However, the lack of diversity in time and weather is a limitation of KITTI. This dataset does not directly provide the ground truth of the test set, and specific requirements are needed for online testing. Therefore, we followed [12] to split the training set into two parts: 3712 samples (training set) and 3769 samples (validation set).

4.1.2. Evaluation metrics

We evaluated the results of both 3D and Bird's-Eye View (BEV) object detection on the KITTI validation set for cars to validate our algorithm. We calculate the corresponding Precision/Recall curves to obtain Average Precision (AP) with threshold values of 0.5 and 0.7. Although KITTI currently adopts AP₄₀ instead of AP₁₁, many existing methods have

したがって、 AP_{11} をAP計算基準として使用する。 AP_{BEV} はBEV物体検出のためのAP、 AP_{3D} は3次元物体検出のためのAPと定義される。KITTIはサンプルの難易度を「簡単」「中程度」「難しい」の3つに分類している。

4.1.3. Baseline

我々の研究は単眼画像からのPseudo-LiDARに基づいているため、単眼画像を入力とする最先端の3D物体検出手法と比較する。MonoDLE[38]の AP_{11} を公式コードに従って求めた¹。

4.2. 実装

4.2.1. 擬似LiDAR生成

KITTIの7481枚の画像に対して、DORNを用いて深度推定を行い、対応する深度マップを得る。深度マップ座標変換と投影変換によって得られる擬似LiDARは、カメラ空間上にある。したがって、KITTIが提供する公式カメラ行列を用いて、視覚点群をVelodyne LiDAR点群に変換する必要がある。

4.2.2. 信頼度サンプリング

局所信頼度を計算する際、物体車両の外部3次元背景が完全にフィルタリングされないように、 $\lambda \alpha$ (重量バランスパラメータ)は5、 $\xi \alpha$ (背景閾値)は0.2とする。大域的信頼度の計算において、長距離点が完全にフィルタリングされないように、 $\lambda \beta$ (大域的バランスパラメータ)は1.5、 $\xi \beta$ (背景閾値)は0.2である。上記のパラメータ構成により、信頼度の高い点はランダムサンプリングにより保持され、重要でない背景点はフィルタリングされる。

4.2.3. 階層的幾何学的特徴抽出

PVRCNNネットワークの入力として、信頼度サンプリングされたPseudo-LiDARが使用される。サンプル数nは2048であるため、この2048個のキーポイントの特徴を最適化した。Kは16に設定される。デュアルトランスクの計算はGPUメモリを消費する。そこで、メモリ使用量を減らすためにキーポイントをサンプリングした。最後に、サンプリングされたキーポイントは、トリリニア補間によってキーポイントの生数に復元された。

4.2.4. Training

PV-RCNNネットワークをベースにネットワークフレームワークを構築し、RT X2080Ti GPUで学習させた。学習率0.01のAdamアルゴリズムを50エポックの学習に使用した。

4.3. Experiment results

4.3.1. 他の手法との比較

表1に、KITTI検証セットにおける我々の手法と最先端手法の検出結果をまとめた。我々の手法は、3つの異なる検出難易度、すなわち、簡単、中程度、難しいにおいて、他の手法を凌駕する。BEVタスクでは、IoU 1/4 0.7(中程度)において、 $AP_{BEV}(\%)$ はAM3Dと比較して約5.4改善した。3D検出タスクでは、IoU 1/4 0.7(中程度)において、 AP_{3D} (単位:%)はMonoFLEXと比較して約5.0改善した。このことは、生成されたPseudo-LiDAR点群の品質と点群の幾何学的構造的特徴的重要性を強く示している。本手法では、車と歩行者を区別しておらず、検出性能は物体の奥行きの推定誤差にのみ依存する。したがって、本手法は自動車の検出性能に応じて歩行者にも有効である。

4.3.2. 回帰パラメータが検出性能に与える影響の比較回帰パラメータ、すなわち、中心

座標(x, y, z)、長さ、幅、高さ(l, w, h)が検出性能に与える影響を調べるために、回帰パラメータの平均二乗誤差(MSE)と、異なる距離における3Dバウンディングボックスの数を数えた。平均二乗誤差は以下の式で求められる：

$$\begin{cases} MSE_{xyz} = \frac{1}{3} ((x_g - x_p)^2 + (y_g - y_p)^2 + (z_g - z_p)^2) \\ MSE_{lwh} = \frac{1}{3} ((l_g - l_p)^2 + (w_g - w_p)^2 + (h_g - h_p)^2) \\ MSE = \alpha MSE_{xyz} + \beta MSE_{lwh} \end{cases} \quad (18)$$

ここで、 (x_g, y_g, z_g) は3次元中心座標、 (l_g, w_g, h_g) はそれぞれ実箱の長さ、幅、高さである。 (x_p, y_p, z_p) はそれぞれ3次元中心座標、 (l_p, w_p, h_p) は予測箱の長さ、幅、高さである。 α と β はそれぞれ20と10の均衡係数である。図4(a)と(b)は、我々の手法が、異なる距離において、ほぼ最小(MSE)と最大数のバウンディングボックスを得ることを示している。いくつかの方法はMSEが50~70mと小さいが、長距離の物体をほとんど検出できない。したがって、回帰パラメータを小さく(MSE)する方法が、より良い検出性能を得ることができることがわかる。

4.4. Ablation studies

4.4.1. 信頼度サンプリング

提案する点群信頼度サンプリング法の妥当性をさらに実証するために、信頼度サンプリングや階層的特徴抽出モジュールを用いないアプローチをベースラインとした。表2にアブレーション実験の結果を示す。信頼度サンプリングモジュールを用いたモデルは、3つのレベルでベースラインと比較して、 AP_{3D} (単位:%)値が2.0、1.4、0.5改善した。図5は、信頼度サンプリング前後の検出結果であり、近いシーンでの検出結果に大きな差はない。しかし、長距離シーンで信頼度サンプリングを実行すると、検出結果が向上する。可視化結果と検出指標から、提案手法が有望な効果を達成していることを証明することができる。

4.4.2. 階層的幾何学的特徴抽出

表3では、Confidence Sampling Module (CSM)とHierarchical Geometric Feature Extraction Module (HGFEM)の検出性能をベースラインと比較して検証している。いずれの場合も、3次元物体検出の性能は向上している。点群の幾何学的特徴を強調することは、単眼Pseudo-LiDAR 3D物体検出のための有望なソリューションである。

4.4.3. デュアル変換器の構造

3次元物体検出の性能に対する変換器構造の影響をさらに検証するために、3つの異なる変換器構造、すなわち、(1)並列構造を持つ二重変換器(並列TR)、(2)カスケード構造を持つ二重変換器(カスケードTR)、(3)単一変換器構造(単一TR)を比較した。便宜上、HGFEMの局所注意特徴エンコーディングモジュールを削除した。表4から、デュアル変換器の性能はシングル変換器よりも優れており、パラレルTRは中程度のサンプルとハードサンプルで最も良い性能を示すが、イージーサンプルではカスケードTRよりも劣っていることがわかる。したがって、並列構造変換器が検証される。

4.5. 可視化分析

図6に本手法の3次元可視化結果を示す。ほとんどの場合、本手法は長距離シーンにおける3Dバウンディングボックスを正確に予測することができ、生成されたPseudo-LiDAR点群の品質の悪さを効果的に扱うことができる。

¹ <https://github.com/xinzuma/monodle>.

only the results of AP₁₁. Therefore, we use AP₁₁ as our AP calculation criterion. AP_{BEV} is defined as the AP for BEV object detection, and AP_{3D} is the AP for 3D object detection. KITTI classifies the difficulty of the samples into three categories: easy, moderate, and hard.

4.1.3. Baseline

Our work is based on Pseudo-LiDAR from monocular images, so we compare it with the state-of-the-art 3D object detection methods that take monocular images as input. We obtained the AP₁₁ of MonoDLE [38] according to the official code.¹

4.2. Implementation

4.2.1. Pseudo-LiDAR generation

We perform depth estimation on 7481 images from KITTI by using the DORN to obtain the corresponding depth maps. The Pseudo-LiDAR obtained by depth map coordinate transformation and projection transformation is on the camera space. Therefore, we need to convert the visual point cloud to a Velodyne LiDAR point cloud by using the official camera matrix provided by KITTI.

4.2.2. Confidence sampling

When calculating the local confidence, λ_α (weight balance parameter) is 5 and ξ_α (background threshold) is 0.2 to ensure that the external 3D background points of the object vehicle are not completely filtered out. In calculating the global confidence, λ_β (global balance parameter) is 1.5, and ξ_β (background threshold) is 0.2 to ensure that the long-distance points are not completely filtered out. With the above parameter configuration, points with high confidence are retained by random sampling, and unimportant background points are filtered out.

4.2.3. Hierarchical geometric feature extraction

The confidence-sampled Pseudo-LiDAR is used as input for the PV-RCNN network. The number of samples n is 2048, so we optimized the features of these 2048 keypoints. K is set to 16. The dual transformer computation consumes GPU memory. Therefore, we sampled the keypoints to reduce memory usage. Finally, the sampled keypoints were restored to the raw number of keypoints by trilinear interpolation.

4.2.4. Training

We constructed our entire network framework based on the PV-RCNN network and trained it on an RTX2080Ti GPU. The Adam algorithm with a learning rate of 0.01 was used for training in 50 epochs.

4.3. Experiment results

4.3.1. Comparison with other methods

In Table 1, we summarize the detection results of our method and state-of-the-art methods on the KITTI validation set for cars. Our method outperforms the other methods on three different detection difficulty levels, i.e., easy, moderate and hard. For the BEV task, AP_{BEV} (in %) is improved by approximately 5.4 compared to AM3D at IoU = 0.7 (moderate). For the 3D detection task, AP_{3D} (in %) is improved by approximately 5.0 compared to MonoFLEX at IoU = 0.7 (moderate). This strongly demonstrates the importance of the quality of the generated Pseudo-LiDAR point clouds and the geometric structural features of the point cloud. Our method does not distinguish cars and pedestrians, and the performance of detection only depends on the estimation error of the object depth. Therefore, our method is also effective for pedestrians according to the detection performance for cars.

4.3.2. Comparing effect of regression parameters on detection performance

To investigate the effect of the regression parameters, i.e., center

coordinates (x, y, z), length, width and height (l, w, h) of the 3D bounding box on detection performance, we counted the Mean Square Error (MSE) of the regression parameters and the number of 3D bounding boxes at different distances. The mean square error is obtained by the following formula:

$$\begin{cases} MSE_{xyz} = \frac{1}{3} \left((x_g - x_p)^2 + (y_g - y_p)^2 + (z_g - z_p)^2 \right) \\ MSE_{lwh} = \frac{1}{3} \left((l_g - l_p)^2 + (w_g - w_p)^2 + (h_g - h_p)^2 \right) \\ MSE = \alpha MSE_{xyz} + \beta MSE_{lwh} \end{cases} \quad (18)$$

Here, (x_g, y_g, z_g) are the 3D center coordinates, and (l_g, w_g, h_g) are the length, width and height of the real boxes, respectively. (x_p, y_p, z_p) are the 3D center coordinates, and (l_p, w_p, h_p) are the length, width and height of the predicted boxes, respectively. α and β are equilibrium coefficients of 20 and 10, respectively.

Figs. 4 (a) and (b) show that our method obtains almost the minimum (MSE) and the maximum number of bounding boxes at different distances. Although several methods have a small MSE of between 50 and 70 m, they can barely detect objects at long distances. Therefore, we know that a method with smaller (MSE) of the regression parameters can obtain better detection performance.

4.4. Ablation studies

4.4.1. Confidence sampling

To further demonstrate the reasonableness of the proposed point cloud confidence sampling method, we took an approach without confidence sampling or hierarchical feature extraction modules as the baseline. Table 2 lists the results of the ablation experiments. The model with a confidence sampling module improves by 2.0, 1.4 and 0.5 in the AP_{3D} (in %) values compared to the baseline on the three levels. Fig. 5 shows the detection results before and after confidence sampling, and there is not much difference between the detection results on the close scenes. However, the detection results improve when confidence sampling is executed on the long-distance scenes. From the visualization results and detection metrics, we can prove that our proposed method achieves a promising effect.

4.4.2. Hierarchical geometric feature extraction

In Table 3, we verify the detection performance of the Confidence Sampling Module (CSM) and the Hierarchical Geometric Feature Extraction Module (HGFEM) compared with the baseline. In both cases, the performance of 3D object detection is improved. Enhancing the geometric features of point clouds is a promising solution for monocular Pseudo-LiDAR 3D object detection.

4.4.3. Dual transformer structure

To further verify the effect of transformer structure on the performance of 3D object detection, we compared three different transformer structures, i.e., (1) dual transformer with a parallel structure (parallel TR), (2) dual transformer with a cascade structure (cascade TR) and (3) single transformer structure (single TR). For convenience, we removed the local attention feature encoding module in the HGFEM. In Table 4, we can see that the performance of the dual transformer is better than that of the single transformer, and the parallel TR performs best on moderate and hard samples but worse than the cascade TR on easy samples. Therefore, the parallel structure transformer is verified.

4.5. Visualization analysis

Fig. 6 shows the 3D visualization results of our methods. In most cases, our method can accurately predict the 3D bounding boxes in long-distance scenes and effectively handle the poor quality of the generated Pseudo-LiDAR point clouds. However, our approach has limitations since

¹ <https://github.com/xinzuma/monodule>.

Table 1

KITTI検証セットでの性能比較。

Method	AP _{BEV} /AP _{3D} (in %), IoU = 0.5			AP _{BEV} /AP _{3D} (in %), IoU = 0.7		
	Easy	Moderate	Hard	Easy	Moderate	Hard
MonoGRNet [13]	-/50.51	-/36.97	-/30.82	-/13.88	-/10.19	-/7.62
M3D-RPN [14]	55.37/48.96	42.49/39.57	35.29/33.01	25.94/20.27	21.18/17.06	17.09/15.21
Pseudo-LiDAR [15]	70.8/66.3	49.4/42.3	42.7/38.5	40.6/28.2	26.3/18.5	22.9/16.4
AM3D [28]	72.64/68.86	51.82/49.19	44.21/42.24	43.75/32.23	28.39/21.09	23.87/17.26
D4LCN [37]	-	-	-	-/26.97	-/21.71	-/18.22
MonoDLE [38]	-	-	-	-/23.75	-/20.71	-/18.00
MonoFLEX [39]	-	-	-	-/28.17	-/21.92	-/19.07
Ours	75.41/71.46	56.59/53.5	50.96/46.2	47.93/39.21	33.83/26.87	28.40/22.60

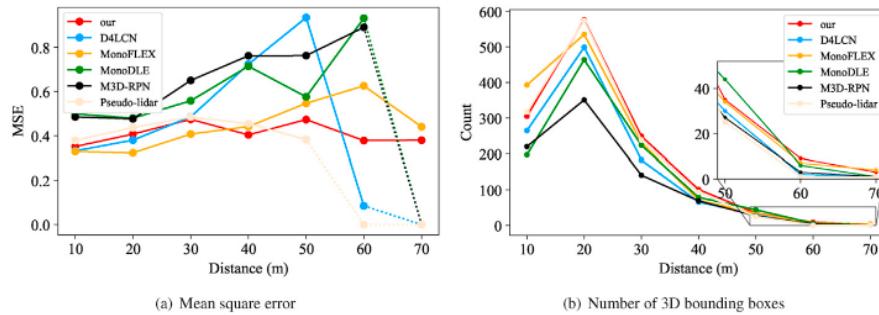


図4. 回帰パラメータの平均二乗誤差(左)と3次元バウンディングボックスの数(右)。(点線はその距離内に物体が検出されなかったことを示す)。

Table 2

信頼度サンプリングのためのアプレーション研究。ローカル信頼度モジュール(pLCM)とグローバル信頼度モジュール(pGCM)を使用した場合の効果を示す。

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
baseline	35.24	24.55	20.75
+LCM	36.10	25.22	20.88
+GCM	36.72	25.30	21.06
+LCM + GCM	37.28	25.92	21.28

Table 3

階層的な幾何学的特徴抽出を用いたアプレーション研究。ベースライン(pHGFEM)と信頼度サンプリングモジュール(pCSM p HGFEM)で階層的幾何学的特徴抽出モジュールを使用した場合の効果を示す。

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
baseline	35.24	24.55	20.75
+HGFEM	37.23	25.45	21.16
+CSM + HGFEM	39.21	26.87	22.60

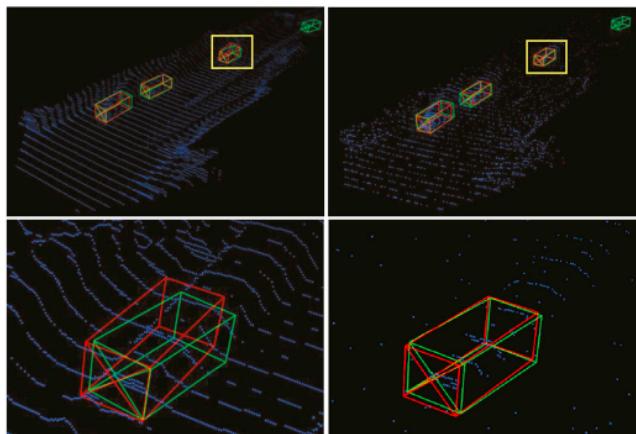


図5 信頼度サンプリング前(左)と後(右)の3次元検出結果。シーン全体の3D検出結果(上)と、黄色枠内の3D検出結果の局所拡大(下)。

しかし、生成されたPseudo-LiDARは、画像で切り捨てられた物体、例えば黄色のボックス内の物体の完全な形状を持っていないため、我々のアプローチには限界がある。今後、これに取り組む予定である。

Table 4

デュアルトランスク構造によるアプレーション研究。

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
単一TRカス	34.26	25.58	20.87
ケードTR並	36.81	25.93	20.86
列TR	36.61	26.00	22.34

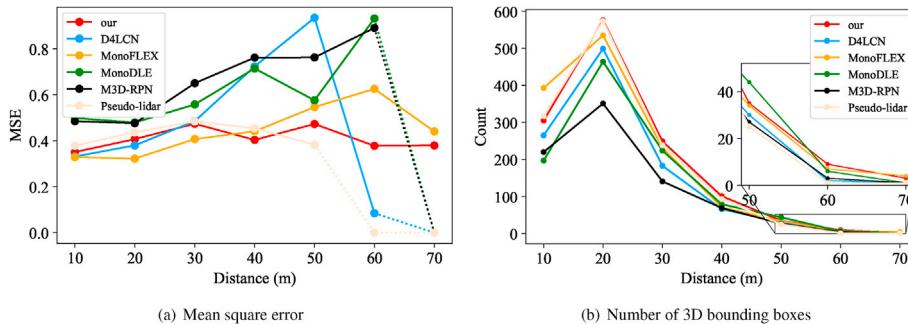
5. Conclusion

本論文では、信頼度サンプリング戦略と階層的な幾何学的特徴抽出モジュールを設計することで、新しい3次元物体検出法を提示した。LiDARを用いた3次元物体検出手法であるPV-RCNNを最大限に活用し、単眼3次元物体検出のための新たなベンチマークを得た。提案したPseudo-LiDAR信頼度サンプリング戦略を点群処理に適用し、大きな奥行き推定誤差の問題に対処し、Pseudo-LiDAR点群の品質を向上させた。提案する階層的幾何学的特徴抽出モジュールは、点群の特徴を抽出するために注意と変換器を用いることで、検出モデルをより幾何特徴に集中させることができる。

Table 1

Performance comparison on KITTI validation set.

Method	AP _{BEV} /AP _{3D} (in %), IoU = 0.5			AP _{BEV} /AP _{3D} (in %), IoU = 0.7		
	Easy	Moderate	Hard	Easy	Moderate	Hard
MonoGRNet [13]	-/50.51	-/36.97	-/30.82	-/13.88	-/10.19	-/7.62
M3D-RPN [14]	55.37/48.96	42.49/39.57	35.29/33.01	25.94/20.27	21.18/17.06	17.09/15.21
Pseudo-LiDAR [15]	70.8/66.3	49.4/42.3	42.7/38.5	40.6/28.2	26.3/18.5	22.9/16.4
AM3D [28]	72.64/68.86	51.82/49.19	44.21/42.24	43.75/32.23	28.39/21.09	23.87/17.26
D4LCN [37]	-	-	-	-/26.97	-/21.71	-/18.22
MonoDLE [38]	-	-	-	-/23.75	-/20.71	-/18.00
MonoFLEX [39]	-	-	-	-/28.17	-/21.92	-/19.07
Ours	75.41/71.46	56.59/53.5	50.96/46.2	47.93/39.21	33.83/26.87	28.40/22.60

**Fig. 4.** Mean square error of regression parameters (left) and number of 3D bounding boxes (right). (Dotted line indicates that no object was detected within that distance.)**Table 2**

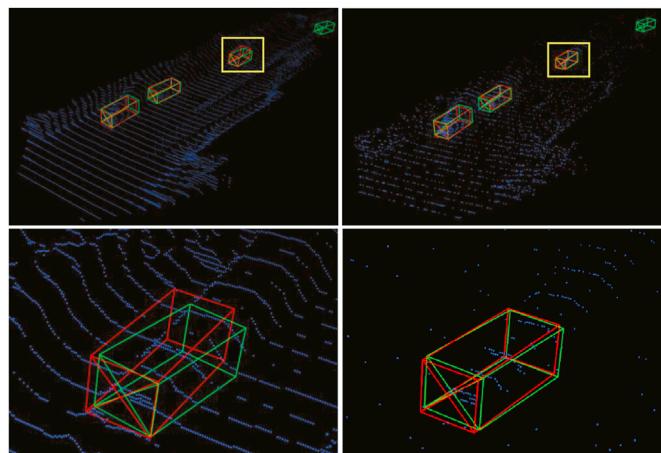
Ablation study for confidence sampling. Effects of using Local Confidence Module (+LCM) and Global Confidence Module (+GCM) are shown.

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
baseline	35.24	24.55	20.75
+LCM	36.10	25.22	20.88
+GCM	36.72	25.30	21.06
+LCM + GCM	37.28	25.92	21.28

Table 3

Ablation study with hierarchical geometric feature extraction. Effects of using Hierarchical Geometric Feature Extraction Module in baseline (+HGFEM) and in Confidence Sampling Module (+CSM + HGFEM) are shown.

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
baseline	35.24	24.55	20.75
+HGFEM	37.23	25.45	21.16
+CSM + HGFEM	39.21	26.87	22.60

**Fig. 5.** 3D detection results before (left) and after (right) confidence sampling. 3D detection results for entire scene (top) and for local enlargement of 3D detection results in yellow box (bottom).

the generated Pseudo-LiDAR does not have the full geometry of the vehicle for the object being truncated in the image, e.g., the object in the yellow box. We will address this in the future.

Table 4

Ablation study with dual transformer structure.

Method	IoU = 0.7, AP _{3D} (in %)		
	Easy	Moderate	Hard
single TR	34.26	25.58	20.87
cascade TR	36.81	25.93	20.86
parallel TR	36.61	26.00	22.34

5. Conclusion

In this paper, we presented a novel 3D object detection method by designing a confidence sampling strategy and a hierarchical geometric feature extraction module. We obtained a new benchmark for monocular 3D object detection by taking full advantage of the LiDAR-based 3D object detection method, i.e., PV-RCNN. The proposed Pseudo-LiDAR confidence sampling strategy was applied to process point clouds, addressing the problem of large depth estimation errors and improving the quality of Pseudo-LiDAR point clouds. The proposed hierarchical geometric feature extraction module uses attention and a transformer to extract the features of point clouds, and this can make the detection model more focused on the geometric features. Experiments verified that the detection performance of our method achieves significant advantages on KITTI (car) on easy, moderate and hard samples. However, there is

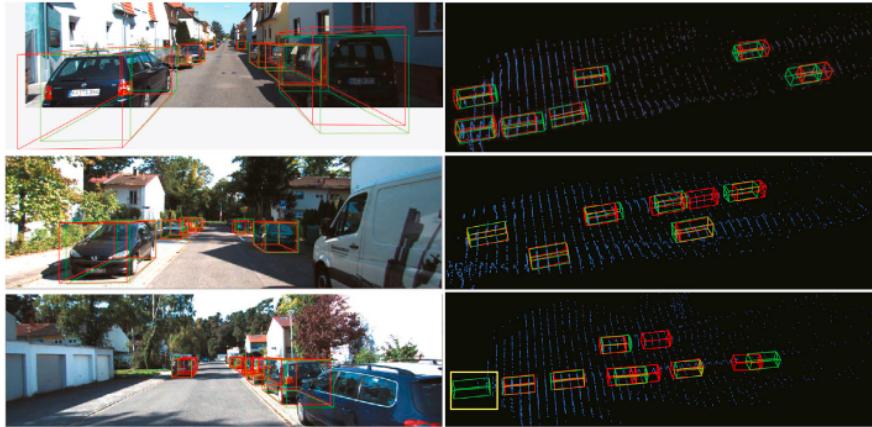


図6. RGB画像(左)とPseudo-LiDAR(右)の3D検出結果。緑色のボックス: グランドトゥルース。赤枠: 予測値。

実験により、本手法の検出性能が、KITTI(車)において、簡単なサンプル、中程度のサンプル、難しいサンプルで有意な優位性を達成することが検証された。しかし、単眼の奥行き推定情報の精度を向上させる余地はまだある。

競合する利益の宣言

原稿に利益相反がないことを宣言する。

謝辞を述べる

本研究は、中国国家重点研究開発計画(2020YFB1807500)、中国国家自然科学基金(62072360、62001357、62172438、61901367)、陝西省重点研究開発計画(2021ZDLGY02-09、2023-GHZD-44、2023-ZDLGY-54)、中国広東省自然科学基金(2022A1515010988)、西安科学技術計画人工知能重点プロジェクト(2022JH-RGZN-0003、2022JH-RGZN-0103、2022JH-CLCJ-0053)、西安科学技術計画(20RGZN0005)および西安大学杭州研究所の概念実証基金(GNYZ2023QC0201)。

References

- [1] T. Qiu, Z. Zhao, T. Zhang, C. Chen, C.P. Chen, Underwater internet of things in smart ocean: system architecture and open issues, *IEEE Trans. Ind. Inf.* 16 (7) (2019) 4297–4307.
- [2] W. Sun, S. Lei, L. Wang, Z. Liu, Y. Zhang, Adaptive federated learning and digital twin for industrial internet of things, *IEEE Trans. Ind. Inf.* 17 (8) (2020) 5605–5614.
- [3] C. Wang, C. Chen, Q. Pei, Z. Jiang, S. Xu, An information centric in-network caching scheme for 5g-enabled internet of connected vehicles, *IEEE Trans. Mobile Comput.* 22 (6) (2023) 3137–3150.
- [4] W. Sun, P. Wang, N. Xu, G. Wang, Y. Zhang, Dynamic digital twin and distributed incentives for resource allocation in aerial-assisted internet of vehicles, *IEEE Internet Things J.* 9 (8) (2022) 5839–5852.
- [5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [6] W. Sun, N. Xu, L. Wang, H. Zhang, Y. Zhang, Dynamic Digital Twin and Federated Learning with Incentives for Air-Ground Networks, *IEEE Trans. Netw. Sci. Eng.* 9 (1) (2022) 321–333.
- [7] C. Chen, L. Liu, S. Wan, X. Hui, Q. Pei, Data dissemination for industry 4.0 applications in internet of vehicles based on short-term traffic prediction, *ACM Trans. Internet Technol.* 22 (1) (2021) 1–18.
- [8] X. Chen, S. Leng, J. He, L. Zhou, Deep learning based intelligent inter-vehicle distance control for 6g-enabled cooperative autonomous driving, *IEEE Internet Things J.* 8 (20) (2021) 15180–15190.
- [9] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
- [10] A. Simonelli, S.R. Bulo, L. Porzi, M. López-Antequera, P. Kuntschieder, Disentangling monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1991–1999.
- [11] B. Xu, Z. Chen, Multi-level fusion based 3d object detection from monocular images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2345–2353.
- [12] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, R. Urtasun, Monocular 3d object detection for autonomous driving, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2147–2156.
- [13] Z. Qin, J. Wang, Y. Lu, Monognet: a geometric reasoning network for monocular 3d object localization, *Proc. AAAI Conf. Artif. Intell.* 33 (2019) 8851–8858.
- [14] G. Brazil, X. Liu, M3d-rpn: monocular 3d region proposal network for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9287–9296.
- [15] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, K.Q. Weinberger, Pseudo-lidar from visual depth estimation: bridging the gap in 3d object detection for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8445–8453.
- [16] C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas, Frustum pointnets for 3d object detection from rgbd data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 918–927.
- [17] C. Chen, Y. Zeng, H. Li, Y. Liu, S. Wan, A multi-hop task offloading decision model in m2c-enabled internet of vehicles, *IEEE Internet Things J.* 10 (4) (2022) 3215–3230.
- [18] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: point-voxel feature set abstraction for 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10529–10538.
- [19] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
- [20] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.
- [21] S. Shi, X. Wang, H. Li, Pointrenn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 770–779.
- [22] C. R. Qi, L. Yi, H. Su, L. J. Guibas, Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, *arXiv preprint arXiv:1706.02413*.
- [23] Y. Yan, Y. Mao, B. Li, Second: sparsely embedded convolutional detection, *Sensors* 18 (10) (2018) 3337.
- [24] Y. Chen, S. Liu, X. Shen, J. Jia, Fast point r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9775–9784.
- [25] C. Yan, E. Salman, Mono3d: open source cell library for monolithic 3-d integrated circuits, *IEEE Trans Circ Syst I: Regul. Pap.* 65 (3) (2017) 1075–1085.
- [26] F. Chatbot, M. Chaouch, J. Rabarisso, C. Teuliére, T. Chateau, Deep manta: a coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2040–2049.
- [27] F. Manhardt, W. Kehl, A. Gaidon, Roi-10d: monocular lifting of 2d detection to 6d pose and metric shape, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2069–2078.
- [28] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, X. Fan, Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6851–6860.
- [29] J. M. U. Vianney, S. Aich, B. Liu, Refinedmpl: Refined Monocular Pseudolidar for 3d Object Detection in Autonomous Driving, *arXiv preprint arXiv:1911.09712*.
- [30] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.
- [31] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, A. Markham, Randlanet, Efficient semantic segmentation of large-scale point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11108–11117.
- [32] L. Wang, Y. Huang, Y. Hou, S. Zhang, J. Shan, Graph attention convolution for point cloud semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10296–10305.



Fig. 6. 3D detection results for RGB image (left) and Pseudo-LiDAR (right). Green boxes: ground truth. Red boxes: predicted.

still room for improving the accuracy of monocular depth estimation information.

Declaration of competing interest

We declare that we have no conflict of interests in our manuscript.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2020YFB1807500), the National Natural Science Foundation of China (62072360, 62001357, 62172438, 61901367), the key research and development plan of Shaanxi province (2021ZDLGY02-09, 2023-GHZD-44, 2023-ZDLGY-54), the Natural Science Foundation of Guangdong Province of China (2022A1515010988), Key Project on Artificial Intelligence of Xi'an Science and Technology Plan (2022JH-RGZN-0003, 2022JH-RGZN-0103, 2022JH-CLCJ-0053), Xi'an Science and Technology Plan (20RGZN0005) and the Proof-of-concept fund from Hangzhou Research Institute of Xidian University (GNYZ2023QC0201).

References

- [1] T. Qiu, Z. Zhao, T. Zhang, C. Chen, C.P. Chen, Underwater internet of things in smart ocean: system architecture and open issues, *IEEE Trans. Ind. Inf.* 16 (7) (2019) 4297–4307.
- [2] W. Sun, S. Lei, L. Wang, Z. Liu, Y. Zhang, Adaptive federated learning and digital twin for industrial internet of things, *IEEE Trans. Ind. Inf.* 17 (8) (2020) 5605–5614.
- [3] C. Wang, C. Chen, Q. Pei, Z. Jiang, S. Xu, An information centric in-network caching scheme for 5g-enabled internet of connected vehicles, *IEEE Trans. Mobile Comput.* 22 (6) (2023) 3137–3150.
- [4] W. Sun, P. Wang, N. Xu, G. Wang, Y. Zhang, Dynamic digital twin and distributed incentives for resource allocation in aerial-assisted internet of vehicles, *IEEE Internet Things J.* 9 (8) (2022) 5839–5852.
- [5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [6] W. Sun, N. Xu, L. Wang, H. Zhang, Y. Zhang, Dynamic Digital Twin and Federated Learning with Incentives for Air-Ground Networks, *IEEE Trans. Netw. Sci. Eng.* 9 (1) (2022) 321–333.
- [7] C. Chen, L. Liu, S. Wan, X. Hui, Q. Pei, Data dissemination for industry 4.0 applications in internet of vehicles based on short-term traffic prediction, *ACM Trans. Internet Technol.* 22 (1) (2021) 1–18.
- [8] X. Chen, S. Leng, J. He, L. Zhou, Deep learning based intelligent inter-vehicle distance control for 6g-enabled cooperative autonomous driving, *IEEE Internet Things J.* 8 (20) (2021) 15180–15190.
- [9] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
- [10] A. Simonelli, S.R. Bulo, L. Porzi, M. López-Antequera, P. Kortschieder, Disentangling monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1991–1999.
- [11] B. Xu, Z. Chen, Multi-level fusion based 3d object detection from monocular images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2345–2353.
- [12] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, R. Urtasun, Monocular 3d object detection for autonomous driving, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2147–2156.
- [13] Z. Qin, J. Wang, Y. Lu, Monognnet: a geometri reasoning network for monocular 3d object localization, *Proc. AAAI Conf. Artif. Intell.* 33 (2019) 8851–8858.
- [14] G. Brazil, X. Liu, M3d-rpn: monocular 3d region proposal network for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9287–9296.
- [15] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, K.Q. Weinberger, Pseudo-lidar from visual depth estimation: bridging the gap in 3d object detection for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8445–8453.
- [16] C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas, Frustum pointnets for 3d object detection from rgb-d data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 918–927.
- [17] C. Chen, Y. Zeng, H. Li, Y. Liu, S. Wan, A multi-hop task offloading decision model in mec-enabled internet of vehicles, *IEEE Internet Things J.* 10 (4) (2022) 3215–3230.
- [18] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: point-voxel feature set abstraction for 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10529–10538.
- [19] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
- [20] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.
- [21] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 770–779.
- [22] C. R. Qi, L. Yi, H. Su, L. J. Guibas, Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, arXiv preprint arXiv:1706.02413.
- [23] Y. Yan, Y. Mao, B. Li, Second: sparsely embedded convolutional detection, *Sensors* 18 (10) (2018) 3337.
- [24] Y. Chen, S. Liu, X. Shen, J. Jia, Fast point r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9775–9784.
- [25] C. Yan, E. Salman, Mono3d: open source cell library for monolithic 3-d integrated circuits, *IEEE Trans. Circ. Syst. I: Regul. Pap.* 65 (3) (2017) 1075–1085.
- [26] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliére, T. Chateau, Deep manta: a coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2040–2049.
- [27] F. Manhardt, W. Kehl, A. Gaidon, Roi-10d: monocular lifting of 2d detection to 6d pose and metric shape, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2069–2078.
- [28] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, X. Fan, Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6851–6860.
- [29] J. M. U. Vianney, S. Aich, B. Liu, Refinedmpl: Refined Monocular Pseudolidar for 3d Object Detection in Autonomous Driving, arXiv preprint arXiv:1911.09712.
- [30] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.
- [31] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, A. Markham, Randlanet, Efficient semantic segmentation of large-scale point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11108–11117.
- [32] L. Wang, Y. Huang, Y. Hou, S. Zhang, J. Shan, Graph attention convolution for point cloud semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10296–10305.

- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv: 1810.04805.
- [35] H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10076–10085.
- [36] H. Zhao, L. Jiang, J. Jia, P. Torr, V. Koltun, Point Transformer, arXiv preprint arXiv: 2012.09164.
- [37] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, P. Luo, Learning depth-guided convolutions for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 1000–1001.
- [38] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, W. Ouyang, Delving into localization errors for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4721–4730.
- [39] Y. Zhang, J. Lu, J. Zhou, Objects are different: flexible monocular 3ds object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3289–3298.

- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv: 1810.04805.
- [35] H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10076–10085.
- [36] H. Zhao, L. Jiang, J. Jia, P. Torr, V. Koltun, Point Transformer, arXiv preprint arXiv: 2012.09164.
- [37] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, P. Luo, Learning depth-guided convolutions for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 1000–1001.
- [38] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, W. Ouyang, Delving into localization errors for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4721–4730.
- [39] Y. Zhang, J. Lu, J. Zhou, Objects are different: flexible monocular 3ds object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3289–3298.